AUTOMATIC INTONATION RECOGNITION OF SINHALA LANGUAGE TO DETECT SPEECH IMPAIRED IN YOUNG CHILDREN

Chathurindu Wickramaarachchi
Faculty of Computing
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
it17166720@my.sliit.lk

Veerandi Kulasekara
Faculty of Computing
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
veerandi.k@sliit.lk

Koliya Pulasinghe
Faculty of Computing
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
koliya.p@sliit.lk

Vijani Piyawardana
Faculty of Computing
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
vijani.p@sliit.lk

Abstract— **Speech recognition is a better approach to extract, identify and analyze specific features in the voice. Further, this can be utilized to analyze tone, accent and word behaviors such as whether the speaker has made a statement, question or command. This approach comes with a system to recognize Sinhala speech of children between 1 to 6 years based on intonation using Machine Learning. The model has been trained using Neural Network by extracting MFCC features, chroma and Mel based on fundamental frequency(f0). The final system can record, analyze and generate a report on the word behavior of the child which can be used by the speech pathologists to track the improvement of the child. Due to the unavailability of a proper database on Sinhala utterances of children, data has been collected manually in a real environment. Hence, data has been filtered and labelled accordingly. Finally, in the training phase, the system has 85% accuracy and overall accuracy of 90% in the testing phase. With the gap of Sinhala related intonation recognition of children, this approach can be further utilized in other research areas as IoT, smart devices and, speech and voice recognition by other researchers.**

Keywords-Machine Learning; Speech recognition; Sinhala Intonation; MLPClassifier

I. INTRODUCTION

Speech recognition is one of the premier domains in Machine Learning (ML) and pattern recognition technology where Hidden Markov Model (HMM) and Artificial Neural Network (ANN) are applied [1]. Because of the importance of speech, most studies have focused on computer-assisted language learning and applied the domain to learn with special types of speech recognition that have many inaccuracies due to dialect, stress, and noise [2]. To identify the unique features in the Sinhala language, the prosodic features such as pitch, accent and stress should be analyzed. Thus, the intonation of speech recognition is the most comprehensive approach to detect prosodic features.

Dialogue expression, acoustic-prosodic features related to the fundamental frequency(f0), length, and voice quality are proposed and evaluated for automated extraction of paralinguistic information [3]. Hence, using descriptive phonological and phonetic features, the speaker's accent and pronunciation features can be analyzed [4].

Children are mastering their language day by day and parents are their teachers in childhood. They follow and trying to pronounce what others say. In that case, the accent and the tone of a child depend on their family and background. There may be cases like delay in starting speaking or wrong pronunciations. It is critical to develop direct and accurate communication in the spoken language of children to avoid wrong communication and accent. Parents must talk with them and correct them with different activities in childhood. When a child having speech impairments and language disorders such as ASD, Apraxia and Autism, they must be examined by speech pathologists direct for the relevant therapeutic sessions [5]. Identifying intonation is important because the tone of our communication expresses the idea exactly to others and shows correct emotions. Due to various dialects in Sri Lanka, questions can be asked like a statement by changing only the pattern of speaking.

The intonation of children with speech and language impairments has received limited attention, even though deficiencies in similar facets of prosody have been hypothesized to underpin language disability. There may be reasons for that like lack of data on normal intonation of children speak in Sinhala, hard to pin down intonation for testing, as well as the widely held belief that intonation is one of the fields least likely to be affected by developmental speech or language impairments. Duration, rhythmic features, and prominent pitch found in studies of speech of child-directed are thought to be particularly significant in this regard. The hypothesis that language problems arise as a result

of a child's difficulty interpreting prosodic features such as intonation and rhythm is one possible implication of the prosodic/phonological bootstrapping hypothesis.

This system was developed using MLPClassifier and has been created an acoustic model by training to returns correct word behavior with the accuracy of 85%. According to acoustic values, system is going to decide what speaker going to communicate. It may be a statement, command, or a question as illustrating in fig. 1 and the system can use it also for the emotional section.

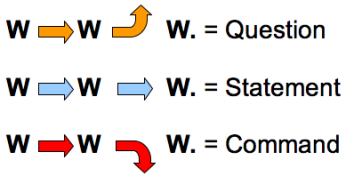

Fig. 1. Word behavior of Sinhala intonation

The final system helps the speech pathologists to have a backup to identify and treat children with any speech or language disorders.

The rest of the paper has been organized as follows: section II discus the existing studies related to intonation and Sinhala speech recognition. The methodology of this study is described in section III, and it describes the system diagram, technologies and methods used to develop the system. Section IV describes testing and accuracy rates of the system and what are future works to do.

II. Related Works

There are many studies conducted on text to speech, speech to text, speaker adaption in the Sinhala language. Also, there are approaches to identify intonation based on f0 in English [6], Japanese [7] languages. Their approaches getting several acoustic features and propose to represent a different type of sentence-final tone. Related to the Japanese language, their approach identifying gentle, carefree, or cheerful, calm, as talking to oneself, consent and disappointed as intonation types. Related works, existing approaches and technologies will be discussed further.

Available approaches of Automatic Speech Recognition (ASR) systems are mainly focused on the middle and young audience. In 2018, Manamperi et al. [8] presents a system to recognize the names and digits of Sinhala songs using speech recognition. Their system is based on HMM and, replace existing Dual Tone Multi-Frequency (DTMF). The system has been developed using Sphinx-4 considering well-prepared documentation, modularity, availability and evolving in real-time. In their approach, it is taking speech signals as inputs and extracting features method of Mel-Frequency Cepstral Coefficients (MFCC) and has been trained the model using HMM. Their system has been done with a word error rate of 11.2% and a sentence error rate of 5.7% respectively by evaluating the performance in a different noisy environment.

In 2005, Asanka Wasala et al. [9] examines major characteristics in Sinhala for detect prosodic features like Syllabification, Stress, Pitch and Intonation related to Phonetics and Phonology. Syllabification research contributes to the discovery of a collection of rules for syllabifying a particular word. The syllabication algorithm had a 99.95% accuracy rate. The letter-to-phoneme mapping was simple due to the phonetic nature of the Sinhala alphabet, but some modifications were needed to generate the appropriate phonetized form of a term. The letter-to-sound conversion algorithm developed in this case, for this reason, has a 98.21% accuracy rate. According to a report on stress assignments, word-level stress is set in the first syllable and long vowels are usually emphasized as well. A critical analysis of the literature was used in their study, as well as further explanation of issues and difficulties experienced by distinguished scholars in linguistics and Sinhala. Their system has used an inventory of Sinhala Phonemes where Sinhala has 40 segmental phonemes, including 14 vowels and 26 consonants, as well as a collection of four prenasalized voiced stops that are unique to Sinhala and Phonotactics in the Sinhala language.

Paul Taylor et al. [10] has done a great approach to the system which explains how to enhance the efficiency of an ASR device by using dialogue context and intonation. The DCIEM map task corpus, a corpus of random task-oriented dialogue speech, was used in their experiments and this corpus was labelled using a dialogue analysis scheme that categorizes each utterance into one of 12 "step forms" like "question-yes/no", "acknowledge" and "instruct". They have used a new method to prosodic phonology and its relationship to conversational structure to incorporate prosodic analysis into the probabilistic context needed for speech recognition. Given that the role of intonation and dialogue that means in their domain is to assist in deciding the most effective language model to identify every auditory communication, it is price questioning whether a lot of simple approaches of simply selecting the language model with the high classification score will succeed.

Based on Natural language processing, in 2018, Gunarathne et al. [11] have the capability of responding to neural language for human use on daily basis by converting speech to text in European languages. The system's recognition output must be evaluated against a corpus of data having tested data set average in 65.84% with a variance of 1000.961 and 31.63797 of standard deviation. However, many approaches having areas related to the proposed approach, there is a gap with the proposed approach. In the next section, the next paragraph will discuss what is the gap between existing approaches and are new features provided by proposed approaches.

The proposed approach is mainly targeting identifying intonation of speech and analyze word behavior, and the domain is targeting children. Existing approaches do not mainly work accurately in identifying children speech but their domain targets young age. Even there is not any proper database or data collection of voice clips of children, speak in Sinhala. Most of the existing systems related to the Sinhala language is text-to-speech or speech-to-text. There is a gap

with finding intonation of the baby to kids based on Sinhala language and identifying speech impairments. Data models have been trained for young to middle-aged persons. It was discussed available methodologies, technologies, algorithms and ways doing related to the proposed approach referring to existing systems.

III. METHODOLOGY

In fig. 2, it has been illustrated 3 initial steps and the major one is preparing data collection by collecting raw data of children speaking in Sinhala, then wav files are converting into numerical data. Then extract features of speeches and send these features and properties into the training module. In that step, it is using a phoneme dictionary and the pronouncing dictionary. In next step, is training data using the acoustic model and using that it is finding correct intonation and output word behaviors.

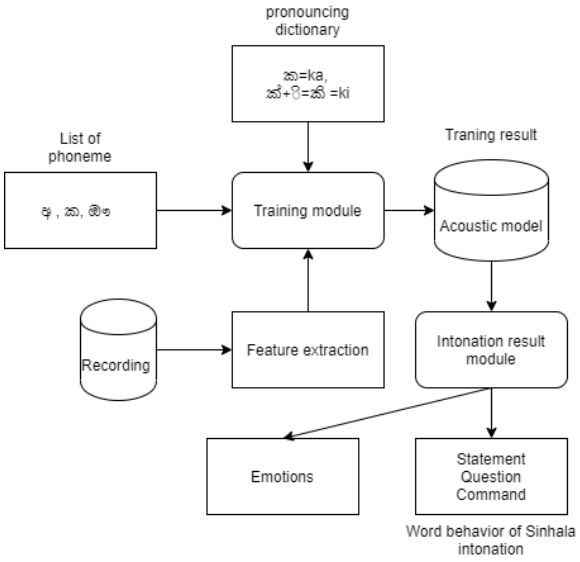

Fig. 2. System diagram of identifying intonation.

Dataset

Sinhala speech dataset is more important than in the proposed ASR system and might have certain qualities to do the experiments. If it is, most open accessible audio datasets have typically been recorded in the sound-proof environment for ignoring noisy sounds and optimize the study's effectiveness. There is not any proper dataset of children speaking in Sinhala. Hence, personal devices were used to record voices in uncontrollable noises. In that case, background noise does not make a different tone for voice clips, but it will be affected by to feature extraction process. Around 1200 voice clips were collected in the age category of two, four, five- and six years including Questions, Statements and Commands. Then labelled according to the child's age, word behavior and emotion by converting to 16000Hz in Mono. When a child expresses his/her feelings using rhythm and pitch, can be identified by the rising and falling of their voice. When a child asking questions his/her pitch tend for rising, in that case, it can be identified as a rising intonation pattern. When a child saying a statement, the pitch falling according to that, but this can be changed due to some facts. There are many cases where children may ask questions in a statement way or some statements expressing like question patterns. It can happen with different patterns, accents in the Sinhala language. Also, there are continuous rising patterns and continuous falling patterns. Table 1 is illustrating some examples of intonation patterns in the Sinhala language.

TABLE I. EXAMPLE OF INTONATION PATTERNS

Types of Sentences	Target Intonation
Questions /Amma ko/ (Where is mom)	Rise
Statement /Mata inne akek/ (I have a sister)	Cont. Rise

Feature Extraction

Every sound is made up of various frequencies, amplitudes, wavelengths, and other attributes - and it is essential to measure these sound characteristics to research them and it is important to extract various sound characteristics of them. Audio features can be mainly classified into three types: Rhythmic features, Spectral features and Temporal features.

Rhythmic features can be mainly used in areas related to music information retrieval. There has been a lot of research conducted in the area of artificial speech emotion detection using rhythmic features [13]. An audio signal sampled over a time frame is described by temporal features. Spectral features which are based on frequencies of audio waves are thought to be the most effective for speech recognition and it is widely used to transform temporal features to frequency counterparts, like the process of reacting to audio signals by human ears [14]. There are several methods to generate frequency domains for spectral features, such as Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC) and Rasta filter.

Multilayer Perceptron (MLP) Classifier

In this research, Sinhala speech was established, considering the effects of tonal coarticulation, intonation and stress by segmenting a tone classification. Automatic syllable segmentation was also established, which divides the training and test utterances into syllable units. The key discriminating features were acoustic features such as fundamental frequency (F0), length, and energy derived from the processing syllable and neighboring syllables. To distinguish these characteristics, a backpropagation approach was used to train a multilayer perceptron (MLP). The suggested scheme was put to the test with 900 test utterances from 10 native Sinhala

speaking children who also delivered the training voice. The proposed approach has an 80% overall accuracy score.

Trimming the edges

As mentioned, that MFCC and its derivates are obtained by splitting an audio signal into small frames ranging from 20ms to 250ms in length and it poses the significant issue of estimating coefficients of undesirable frames. Unnecessary frames are essentially framing that appear at the beginning and end of a sentence but contain no useful details in the domain. These kinds of unnecessary frames are creating when it is starting to record speeches and stopping it after the speaker stopping the sentences. There are a set of methods to removing unnecessary frames, and most the approaches have been used to setting decibel threshold values and anything beyond that calling silence. Then edges can be trimmed by measuring the average decibel value for each frame and if that is less than a certain threshold. Fig. 3 can see how unnecessary frames has been recorded and fig. 4 shows how the audio is after trimming edges.

Fig. 3. Audio file with unnecessary frames.

Fig. 4. Audio file after trimming edges.

The overall process was to move for feature extraction techniques and use feature extraction for the training module using the data collection. Further training the module is using defined Sinhala phoneme dictionary and pronouncing dictionary. The next process was to build the acoustic model and model for intonation result. After identifying intonation properly system can identify word behavior and move for emotion recognition. Finally, the system was tested by tuning parameters to generate a result with a report in high accuracy. The approach requires a minimalized UI to catch user's voice and generate the report. Then it can be used by a speech pathologist to examine children and identifying whether a child is speaking in the correct accent, tone and pitch due to situation and emotions. The next section discusses how testing was done and accuracy rates in age groups.

IV. Testing, Implementation, Results & Discussion

An acoustic feature describes the wave properties of the speech which includes energy-based features, Fourier frequencies, MFCC and similar. Fig. 5, spectrogram visualize the low and high energy of voice clip which is the statement of "/Mata inne ammai, thaththai, nangi/" (I have mother, father and sister) where yellow bands correspond to the frequencies that are strongly present and lowest, or the fundamental frequency (f0) and the higher frequency one is harmonic. In fig.6, it is clear to see where the audio is voiced and where it is unvoiced.

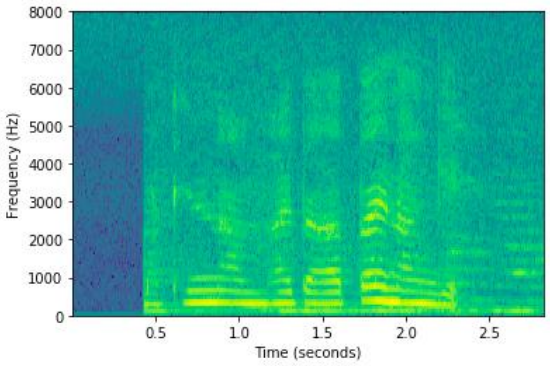

Fig. 5. spectrogram view of high and low energy.

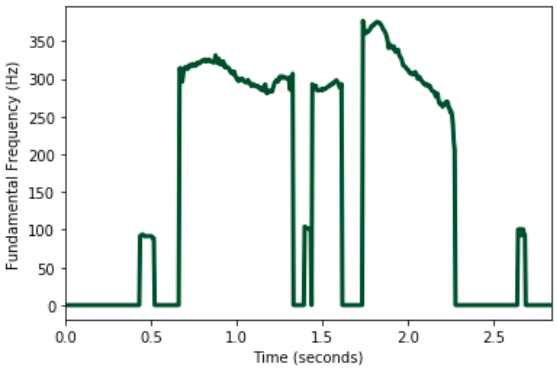

Fig. 6. voice and unvoiced using f0.

Data collection Tuning

The system was trained 20 times in 3 phases: initial step, 50% stage and 95% stage. Initially, around 30 to 70 speech clips were collected from two children and labelled according to statement behavior, age and id of the child. A child spoke in high pitched tone and a different accent, the other child was a little bit calm and talked in a normal accent and vocal tone. When it is training 62% of accuracy was gained, but when testing that model results with 3 children having 10 statements and 9 questions, the system was able to identify 7 statements and ignoring 5 questions as 'not a statement. Then another 30 questions were added from 3 children and got 56% accuracy when training and, the system identified 8 statements out of 10 and ignoring 7 questions as 'not a statement out of 9.

In the 50% stage system was trained with 200 statements from 5 children and 150 questions from 6 children and got an accuracy of 72%. Then test the model with 20 statements and 20 questions of 5 children, the system identified 19 statements correctly and 18 questions correctly. Then additional 50 statements and 50 questions of two children were added by repeating the same statement more than 10 times in different accents and tones, and the same as asking same questions more than 10 times in different tones. After training, the model got 78% accuracy and the model was tested using 25 statements and 25 questions of 3 children, and the system was able to correctly identify 25 statements and 23 questions.

In 95% stage, audio clips of 10 children with 300 statements, 300 questions and 300 commands were trained and got an accuracy of 80%. Then the model was tested using 50 statements, 40 questions and 60 commands of 10 children, and the system finally correctly identified 50 statements, 37 questions and 49 commands.

When increasing the data collection size, what could be noted was system can identify statements of children in the age category 3 to 6 years with 99% accuracy, in the age category 1 to 3 years with 90% accuracy and 1 to 6 years with 95% accuracy. Then the system can recognize correct word behavior of questions of children in age category3 to 6 years with 97% accuracy, 1 to 3 years with 85% accuracy and 1 to 6 years with 90% accuracy. Finally, the system can recognize commands of children in age range 3 to 6 years with 98% accuracy, in age range 1 to 3 years with 92% accuracy and in the age range 1 to 6 years with 94% accuracy. Above mentioned all accuracies were gained when tunning the range and the number of audio clips in data collection.

TABLE II. OVERALL TEST ACCURACY AFTER TUNNING DATA COLLECTION

	1-3 YEARS	3-6 YEARS	1-6 YEARS
STATEMENT	90%	99%	95%
QUESTION	85%	97%	90%
COMMAND	92%	98%	94%

Parameter Tuning

Since there is not having standard value for batch size and epochs, experimented with different values by increasing and decreasing for improving the performance of the algorithm. Finally, the model was trained by giving 0.01 of alpha value, 256 of batch size, 300 of hidden layers in adaptive learning rate as model parameters.

V. CONCLUSION

Communication is the major key of children where the mastering their mother language day by day. Correct intonation in the Sinhala language is a very important factor where it is not only depending on pitch version but, always contain an accompanying version in other prosodic traits such as range of pitch, tone, tempo, loudness, rhythm. Correct intonation of a child describes the difference of statement, command or question and it is affected to the correct emotion expressing of a child. Rising, Falling and peaking (how pitch behaves with time) can be used to describe intonation patterns. There are different accents, the vocabulary used in Sri Lanka and those are depending on areas, culture or parents' backgrounds. Also, some children are getting more time for mastering their speaking ability or vocabulary. It is depending on different facts and there may be cases of some children having speech impairments or Autism cases. In that case, speech therapists must examine them and start to treat them. Focusing on the last point mentioned above, this approach was able to develop a tool for identifying correct intonation of a child in 1 to 6 years speaking in Sinhala to detect whether addressing correct word behavior and examine these points, to make the second opinion for getting strong decision for speech pathologists.

The Data collection is separated into 3 sections: Statement, Command and Question by labelling according to the child, age, word behavior, emotion. Then filtered audio and converted them into 16000Hz wav files in Mono. Then extracted features and properties additionally using phoneme and pronunciation dictionaries. Then trained the model and output the correct word behaviors from the intonation module. Final System has an accuracy rate of up to 95% to identify statements, 90% of accuracy to identify questions and an accuracy rate up to 94% to identify commands. A web-Based application was built for the usage of speech pathologists and using UI, speech pathologists can get records of children or upload and predict already recorded audio files, and using exercises therapy room, they can predict child word behavior related to a video or a picture.

This research has more value where speech and voice recognition having the golden market in industry and, especially the gap of identifying correct intonation of children speak in Sinhala, this tool can be developed further and used for related different domains. Also, other researchers can use this model for their domains.

VI. FUTURE WORKS

When the size of the data set is increased, the accuracy got high value and identified word behaviors more correctly with high accuracy. There is a low level of identifying word behaviors for the children in the age range 1 to 3. Need to add more feature to treat separately for that age range by modifying the algorithm. Due to, there is not any proper data set related to the intonation of children speak in Sinhala, there are limitations to get recordings. Our target dataset is to collect 5000 samples to build the model with high accuracy. Also, the back end must analyze each property, word behaviors, records of the patient, comments of doctors properly and generate a proper medical report the from front end.

ACKNOWLEDGMENT

This research was supported by Sri Lanka Institute of Information Technology and the Accelerating Higher Education Expansion and Development (AHEAD) Operation of the Ministry of Higher Education of Sri Lanka funded by the World Bank.

REFERENCES

[1] Balwant A. Sonkamble and D. D. Doye, "An Overview of Speech Recognition System based on the Support Vector Machines," in *Computer and Communication Engineering*, 2008.

[2] Deguo Mu, Wei Sun, Guoliang Xu and Wei Li, "Japanese Pronunciation Evaluation Based on DDNN," in *IEEE Access* , 2020.

[3] A. D. Cristo and Daniel Hirst, "Intonation systems: A survey of twenty languages.," in *Cambridge: Cambridge University Press*, 2001.

[4] Sagisaka, Y., N. Campbell and N. Higuchi, "Computing Prosody: Computational Models for Processing Spontaneous Speech. Springer," in *Springer-Verlag*, New York.

[5] Gayani Rupasinghe, Roshan Abeyweera, Randil Pushpananda and Ruvan Weerasinghe, "A Mobile-Based Alphabet Learning Game To Intervene Dyslexia Among Children," in *IEEE*, Colombo, Sri Lanka, 2021 .

[6] Ken Ross and Mari Ostendorf, "A Dynamical System Model for Recognizing Intonation Patterns," in *EUROSPEECH '95Fourth European Conference on Speech Communication and Technology*, Madrid, Spain, 1995.

[7] Carlos Toshinori Ishi, Nobuaki Minematsu and Keikichi Hirose, "Recognition of Accent and Intonation Types of Japanese using F0 parameters Related to Human Pitch Perception," in *Prosody in Speech Recognition and Understanding*, USA, 2001.

[8] Wageesha Manamperi, Dinesha Karunathilake, Thilini Madhushani, Nimasha Galagedara and Dileeka Dias, "Sinhala Speech Recognition for Interactive Voice Response Systems Accessed Through Mobile Phones," in *Moratuwa Engineering Research Conference (MERCon)*, Colombo, Sri Lanka, 2018.

[9] A Wasala and K Gamage, "Research Report on Phonetics and Phonology of Sinhala," in Language Technology Research Laboratory, *University of Colombo School of Computing*, Sri Lanka 2005.

[10] Paul Taylor, Simon King, Stephen Isard and Helen Wright, "Intonation and dialogue context as constraints for speech recognition," in *Centre for Speech Technology Research, University of Edinburgh*, UK.

[11] W. T. V. L. Gunarathne, T. K Ramasinghe, D. G. J. B. Wimalarathne, BMSH Balasuriya and B. Hettige, "Sinhala Speech to Text Library using Sphinx," in *General sir John Kotelawala Defense University*, Sri Lanka, 2018.

[12] Bill Wells and Sue J. E. Peppe, "Intonation Abilities of Children With Speech and Language Impairments," in *Journal of Speech Language and Hearing Research*, 2003.

[13] Mayank Bhargava and Tim Polzehl, "Improving Automatic Emotion Recognition from speech using Rhythm and Temporal feature," in *School of Electrical Sciences, Indian Institute of Technology Bhubaneswar,Bhubaneswar*, 2013.

[14] K Han, Y He, D Bagchi, E Fosler-Lussier and D. Liang Wang, "DeepNeuralNetworkBasedSpectralFeatureMapping forRobustSpeech Recognition," in *Department of Computer Science and Engineering & Center for Cognitive and Brain Sciences, 2015*.